arXiv:1006.4582v1 [math.ST] 23 Jun 2010

The Poisson Compound Decision Problem Revisited

L. Brown,[*]

University of Pennsylvania; e-mail: lbrown@wharton.upenn.edu

E. Greenshtein

Israel Census Bureau of Statistics; e-mail: eitan.greenshtein@gmail.com

and

Y. Ritov[†]

The Hebrew University of Jerusalem; e-mail: yaacov.ritov@gmail.com

Abstract: The compound decision problem for a vector of independent Poisson random variables with possibly different means has half a century old solution. However, it appears that the classical solution needs smoothing adjustment even when there are many observations and relatively small means such that the empirical distribution is close to its mean. We discuss three such adjustments. We also present another approach that first transforms the problem into the normal compound decision problem.

1. Introduction

In this paper we consider the problem of estimating a vector $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$, based on observations $Y_1, \dots, Y_n$, where $Y_i \sim Po(\lambda_i)$ are independent. The performance of an estimator $\hat{\boldsymbol{\lambda}}$ is evaluated based on the risk

$$E_{\boldsymbol{\lambda}} ||\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}||^2, \tag{1}$$

which corresponds to the loss function

$$L_2(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = \sum (\lambda_i - \hat{\lambda}_i)^2.$$

Empirical Bayes (EB) is a general approach to handle compound decision problems. It was suggested by Robbins, see (1951, 1955); see Copas (1969) and Zhang (2003) for review papers. Suppose we assume that λ_i, $i = 1, \dots, n$ are realizations of i.i.d. Λ_i, $i = 1, \dots, n$, where $\Lambda_i \sim G$, then a natural approach is to use the Bayes procedure:

$$\delta^G = \operatorname*{argmin}_{\delta} E_G(\delta(Y) - \Lambda)^2, \tag{2}$$

[*] Research supported by an NSF grant.
[†] Research supported by an ISF grant.

and estimate $\boldsymbol{\lambda}$ by $\hat{\boldsymbol{\lambda}} = (\delta^G(Y_1), \ldots, \delta^G(Y_n))$. When G is completely unknown, but it is assumed that $\lambda_1, \ldots, \lambda_n$ are i.i.d., then it may be possible to estimate δ^G from the data $Y_1, \ldots, Y_n$, and replace it by some $\hat{\delta}^G$. Optimal frequentist properties of δ^G in the context of the compound decision problem, are described in terms of optimality within the class of simple symmetric decision functions. See the recent paper by Brown and Greenshtein (2009) for a review of the topic. The optimality of the empirical Bayes decision within the larger class of permutational invariant decision functions is shown in Greenshtein and Ritov (2009).

The Bayes procedure δ^G has an especially simple form in the Poisson setup. In this case there is also a simple and straightforward estimator $\hat{\delta}^G$ for δ^G. Denote by P the joint distribution of (Λ, Y), which is induced by G. The Bayes estimator of λ_i given an observation $Y_i = y$, is:

$$\begin{aligned}\delta^G(y) \equiv E(\Lambda_i | Y_i = y) &= \frac{\int \lambda P(Y_i = y | \Lambda_i = \lambda) dG(\lambda)}{\int P(Y_i = y | \Lambda_i = \lambda) dG(\lambda)} \\ &= \frac{(y+1) P_Y(y+1)}{P_Y(y)},\end{aligned} \tag{3}$$

where P_Y is the marginal distribution of Y under P. Given $Y_1, \ldots, Y_n$, we may estimate $P_Y(y)$ trivially by the empirical distribution: $\hat{P}_Y(y) = \#\{i | Y_i = y\}/n$. We obtain the following Empirical Bayes procedure

$$\hat{\delta}^G(y) = \frac{(y+1)\hat{P}_Y(y+1)}{\hat{P}_Y(y)}. \tag{4}$$

This is currently the "default"/"classical" empirical Bayes estimator in the Poisson setup, suggested initially by Robbins (1955). Various theoretical results established in the above mentioned papers and many other papers, imply that as $n \to \infty$, the above procedure will have various optimal properties. This is very plausible, since as $n \to \infty$, $\hat{P}_Y \to P_Y$ and thus $\hat{\delta}^G \to \delta^G$. However, the convergence may be very slow, even in common situationsn as demonstrated in the following example, and one might want to improve the above $\hat{\delta}^G$. This is the main purpose of this work.

Example 1: Consider the case where $n = 500$ and $\lambda_i = 10$, $i = 1, \ldots, 500$. The Bayes risk of δ^G for a distribution/prior G with all its mass concentrated at 10 is, of course, 0. The risk of the naive procedure which estimates λ_i by Y_i, equals the sum of the variances, that is, $10 \times 500 = 5000$. In 100 simulations we obtained an average loss of 4335 for the procedure (4), which is not a compelling improvement over the naive procedure, and very far from the Bayes risk.

We will improve $\hat{\delta}^G$ mainly through "smoothing". A non-trivial improvement is obtained by imposing monotonicity on the estimated decision function. By imposing monotonicity without any further step, the average total loss in the above example in 100 simulations is reduced to 301; by choosing a suitable smoothing parameter ($h = 3$, see Section 2 below) and imposing monotonicity, the average loss is reduced further to 30. Early attempts to improve (4) through smoothing, including imposing monotonicity, may be found in Maritz (1969) and references there.

The rest of the paper is organized as follows. In Section 2 we will suggest adjustments and improvements of $\hat{\delta}^G$. In Section 3 we describe the alternative approach of transforming the Poisson EB problem to a normal EB problem, using a variance stabilizing transformation. In Section 4 we discuss some decision-theoretic background, and in particular we examine loss functions other than the squared loss. In Section 5 we discuss the above mentioned two approaches and compare them in a simulation study. Both approaches involve a choice of a "smoothing-parameter". A choice based on cross-validation is suggested in Section 6. In Section 7 we present an analysis of real data, describing car accidents. Finally, in Section 8 we briefly describe a further third approach, which estimates δ^G using a nonparametric MLE.

2. Adjusting the classical Poisson empirical Bayes estimator

In the Introduction we introduced the Bayes decision function δ^G and its straightforward estimator $\hat{\delta}^G$. Surprisingly, it was found empirically that even for n relatively large, when the empirical distribution is close to its expectation, the estimated decision function should be smoothed. We discuss in this section how this estimator can be improved. The improvement involves three steps, which finally define an adjusted Robbins estimator.

2.1. Step 1

Recall the joint probability space defined on (Y, Λ). We introduce a r.v. $N \sim Po(h)$, where N is independent of Y and Λ. Let $Z = Y + \mathcal{N}$. Consider the suboptimal decision function

$$\delta_{h,1}(z) \equiv E(\Lambda|Z = z) = E(\Lambda + h|Z = z) - h. \tag{5}$$

The above is the optimal decision rule, when obtaining the corrupted observations $Z_i = Y_i + N_i$, $i = 1, \dots, n$ instead of the observations $Y_1, \dots, Y_n$. The "corruption parameter" h is a selected parameter, referred to as "smoothing parameter". In general, we will select smaller h as n becomes larger. See Section 6 for further discussion on the choice of h. Motivated by (5) and reasoning similar to (4), we define $\hat{\delta}_{h,1}$ as:

$$\hat{\delta}_{h,1}(z) = \frac{(z+1)\tilde{P}_Z(z+1)}{\tilde{P}_Z(z)} - h. \tag{6}$$

where the distribution $\tilde{P}_Z(z)$ is defined by

$$\tilde{P}_Z(z) = \sum_{i=0}^{z} \hat{P}_Y(i) \times \exp(-h) \frac{h^{z-i}}{(z-i)!}. \tag{7}$$

Note that $\tilde{P}_Z(z)$ as defined in (7) involves observation of Y through the quantity $\hat{P}_Y(y)$ that appears inside its definition. It is—in general—a much better estimate of $P_Z(z)$ than is the empirical distribution function $\#\{i: \ Z_i = z\}$.

2.2. Step 2.

There is room for considerable improvement of $\delta_{h,1}$. Note that $\delta_{h,1}$ is applied to the randomized observation Z_i. Therefore, the natural next adjustment is Rao-Blackwellization of the estimator. Define

$$\hat{\hat{\delta}}_{h,2}(y) = E_{\mathcal{N}}(\hat{\hat{\delta}}_{h,1}(y + \mathcal{N})), \tag{8}$$

for $\mathcal{N} \sim Po(h)$, which is independent of the observations $Y_i, \ i = 1, \dots, n$. That is,

$$\hat{\hat{\delta}}_{h,2}(y) = e^{-h} \sum_{j=0}^{\infty} \frac{h^j}{j!} \delta_{h,1}(y + j).$$

Note that for a given y, the value of $\hat{\hat{\delta}}_{h,2}(y)$ depends on all of $\hat{P}_Y(0), \hat{P}_Y(1), \dots,$ although mainly on the values in the neighborhood of y.

2.3. Step 3

Finally after applying adjustments 1 and 2 we obtain a decision function which is not necessarily monotone. However, because of the monotone likelihood ratio property of the Poisson model, δ^G is monotone. A final adjustment is to impose monotonicity on the decision function $\hat{\hat{\delta}}_{h,2}$. We do it through applying isotonic regression by the pulling adjacent violators, cf. Robertson, Wright, and Dykstra (1988). Note, the monotonicity is imposed on $\hat{\hat{\delta}}_{h,2}$ confined to the domain $D(Y) \equiv \{y: \ Y_i = y \text{ for some } i = 1, \dots, n\}$. To be more explicit, an estimator is isotonic if

$$y_i, y_j \in D(Y) \text{ and } y_i \le y_j \Rightarrow \delta(y_i) \le \delta(y_j), \tag{9}$$

and $\delta_{h,3}$ is isotonic and satisfies

$$\sum_{i=1}^{n} \big(\hat{\delta}_{h,3}(y_i) - \hat{\hat{\delta}}_{h,2}(y_i)\big)^2 = \min\Big\{\sum_{i=1}^{n} \big(\hat{\delta}(y_i) - \hat{\hat{\delta}}_{h,2}(y_i)\big)^2 : \delta \text{ satisfies } (9)\Big\}.$$

We obtain the final decision function $\hat{\delta}_{h,3}$, after this third step.

In order to simplify notations we denote: $\Delta_h \equiv \hat{\delta}_{h,3}$. This is our adjusted Robbins estimator.

Finally we remark on a curious discontinuity property of Δ_h. The function Δ_h is a random function, which depends on the realization $\boldsymbol{y} = (y_1, \dots, y_n)$. In order to emphasize it we write here $\Delta_{\boldsymbol{y},h} \equiv \Delta_h$. Consider the collection of functions parameterized by h, denoted $\{\Delta_{\boldsymbol{y},h}(y)\}$. It is evident from the definition of (6), that $\Delta_{\boldsymbol{y},h}(y)$ does not (necessarily) converge to $\Delta_{\boldsymbol{y},0}(y)$ as h approaches 0, even for y in the range $y_1, \dots, y_n$. This will happen whenever there is a gap in the range of y. Suppose, for simplicity that $\hat{P}_Y(y) = 0$, while $\hat{P}_Y(y-1), \hat{P}_Y(y+1) > 0$. Then, $\lim_{h \to 0} \hat{\delta}_{h,1}(y-1) = 0$, and $\lim_{h \to 0} h\hat{\delta}_{h,1}(y) = (y+1)\hat{P}_Y(y+1)/\hat{P}_Y(y-1)$. Hence

$$\begin{aligned} \lim_{h \to 0} \hat{\delta}_{h,2}(y-1) &= \lim_{h \to 0} E\Big(\hat{\delta}_{h,1}(y-1+N)\big|y_1, \dots, y_n\Big) \\ &= \lim_{h \to 0} \big((1-h)\hat{\delta}_{h,1}(y-1) + h\hat{\delta}_{h,1}(y)\big) \\ &= (y+1)\hat{P}_Y(y+1)/\hat{P}_Y(y-1), \end{aligned}$$

which is strictly different from $\delta_{0,2}(y) = 0$. Suppose that $\hat{P}_y(y) > 0$ and $\hat{P}_Y(y+j_0) > 0$ for some $j_0 > 1$, but $\hat{P}(y+j) = 0$ for $j = 1, \dots, j_0 - 1$. Then one can check directly from the definition that $\lim_{h \to 0} \hat{\delta}_{h,2} = y + j_0$. Note that in such a situation $\hat{\delta}^G(y) = 0$. Hence $\hat{\delta}_{h,2}(y)$ for small to moderate h seems preferable to $\hat{\delta}^G(y) = \hat{\delta}_{0,2}(y)$ in such gap situations.

This phenomena is reflected in our simulations, Section 5, especially in Table 5.

Another curious feature of our estimator is when applied on $y_{max} = \max\{Y_1, ..., Y_n\}$. It may be checked that: $\hat{\delta}_{h,2}(y_{max}) = (y_{max}+1)h + O(h^2)$. When h is small so that $(y_{max}+1)h \ll y_{max}$, this would introduce a significant bias. Hence, choosing very small h, might be problematic, though this bias is partially corrected through the isotonic regression. An approach to deal with this curiosity could be to treat y_{max} separately, for example decide on the value of the decision function at the point y_{max} through cross-validation as in Section 6, trying a few plausible values.

3. Transforming the data to normality.

The emprical Bayes approach for the analogous normal problem has also been studied for a long time. See the recent papers of Brown and Greenshtein (2009) and of Wenhua and Zhang (2009) and references there. The Poisson problem and the derivation of (4) are simpler and were obtained by Robbins at a very early stage, before the problem of density estimation, used in the normal empirical Bayes procedure, was addressed. In what follows we will consider the obvious modification of the normal method to the Poisson problem.

In the normal problem we observe $Z_i \sim N(M_i, \sigma^2)$, $i = 1, \dots, n$ where $M_1, \dots, M_n$ are i.i.d. random variables sampled from G and the purpose is to estimate $\mu_1, \dots, \mu_n$ the realizations of $M_1, \dots, M_n$. The application of the normal EB procedure has a few simple steps. First we transform the Poisson variables $Y_1, \dots, Y_n$ to the variables $Z_i = 2 * \sqrt{Y_i + q}$. Various recommenations for q are given in the literature, the simpler and most common one is $q = 0$, the choice $q = 0.25$ was recommended by Brown et. al. (2005, 2009). We now treat Z_i's as (approximate) normal variables with variance $\sigma^2 = 1$ and mean

$2 * \sqrt{\lambda_i}$, and estimate their means by $\hat{\mu}_i$, through applying normal empirical Bayes technique; specifically, $\hat{\mu}_i = \delta_{N,h}(Z_i)$, as defined in (11) below. Finally we estimate $\lambda_i = EY_i$, by $\hat{\lambda}_i = \frac{1}{4}\hat{\mu}_i^2$.

We will follow the approach of Brown and Greenshtein (2009). Let

$$g(z) = \int \frac{1}{\sigma}\varphi\Big(\frac{z-\mu}{\sigma}\Big)dG(\mu).$$

It may be shown that the normal Bayes procedure denoted δ_N^G, satisfies:

$$\delta_N^G(z) = z + \sigma^2 \ \frac{g'(z)}{g(z)}. \tag{10}$$

The procedure studied in Greenshtein and Brown (2009), involves an estimation of δ_N^G, by replacing g and g' in (10) by their kernel estimators which are derived through a normal kernel with bandwidth h. Denoting the kernel estimates by $\hat{g}_h$ and $\hat{g}'_h$ we obtain the decision function, $(Z_1, \dots, Z_n) \times z \mapsto R$:

$$\delta_{N,h}(z) = z + \sigma^2 \ \frac{\hat{g}'_h(z)}{\hat{g}_h(z)}. \tag{11}$$

One might expect this approach to work well in setups where λ_i are large, and hence, the normal approximation to $Z_i = \sqrt{Y_i + q}$ is good. In extensive simulations the above approach was found to also work well for configurations with moderate and small values of λ. In many cases it was comparable to the adjusted Poisson EB procedure.

Remark In the paper of Brown and Greenshtein the estimator $\delta_{N,h}$ as defined in (11) was studied. However, just as in the Poisson case, it is natural to impose monotonicity. In the simulations of this paper we are making this adjustment using isotonic regression. Again, the monotonicity is imposed on $\delta_{N,h}$ confined to the range $\{y_1, ..., y_n\}$. We denote the adjusted estimator by

$$\Delta_{N,h}.$$

4. The loss functions.

The estimator $\delta_{N,h}(Z_i) = \hat{\mu}_i$, may be interpreted as an approximation of the nonparametric EB estimator for $\mu_i \equiv 2\sqrt{\lambda_i}$, based on the (transformed) observations Z_i under the loss $L(\boldsymbol{\mu}, \boldsymbol{a}) = ||\boldsymbol{\mu} - \boldsymbol{a}||^2$, for the decision $\boldsymbol{a} = (a_1, \dots, a_n)$. Thus, $\frac{1}{4}\hat{\mu}_i^2$ may be interpreted as the approximation of the empirical Bayes estimator for λ_i, under the loss

$$L_H(\boldsymbol{\lambda}, \boldsymbol{a}) = \sum(\sqrt{\lambda_i} - \sqrt{a_i})^2 = -2\log(1 - D_H^2),$$

where D_H is to the Hellinger distance between the distributions $\prod Po(\lambda_i)$ and $\prod Po(a_i)$.

Some papers that discuss the problem of estimating a vector of Poisson means are Clevenson and Zidek (1975), Johnstone (1984), Johnstone and Lalley (1984). Those and other works suggest that a particularly natural loss function in addition to L_H and L_2, denoted L_{KL} is

$$L_{KL}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = \sum \frac{(\lambda_i - \hat{\lambda}_i)^2}{\lambda_i}.$$

Note, L_{KL} also corresponds to the local Kulback-Leibler distance between the distributions.

The above articles and some other literature, strongly suggest that L_{KL} is the 'most natural' general loss function. From an empirical Bayes perspective, the optimal decisions that correspond to those three loss functions may have more and less similarity, depending on the configuration. For example, when the prior G is concentrated on a point mass, the Bayes procedures corresponding to those 3 loss functions are obviously the same. Since the L_{KL} loss is of a special importance, we will briefly describe how our analysis can be modified to handle it. As in the case of L_2 loss, one may obtain that the Bayes decision under the L_{KL} loss is given for $y \geq 1$ by:

$$\frac{yP_Y(y)}{P_Y(y-1)}.$$

The decision for $y = 0$ denoted $\hat{\lambda}(0)$, is:

$$\begin{aligned} \hat{\lambda}(0) &= \arg\min_a \int \frac{(\lambda - a)^2}{\lambda} e^{-\lambda} dG(\lambda) \\ &= \frac{\int e^{-\lambda} dG(\lambda)}{\int \lambda^{-1} e^{-\lambda} dG(\lambda)}. \end{aligned}$$

In particular, $\hat{\lambda}(0) = 0$ if G gives a positive probability to any neighborhood of 0.

The decision for $y \geq 1$ may be estimated as in (4) together with the three adjustments suggested in Section 2, along the same lines. However, we still need to approximate the Bayes decision $\hat{\lambda}(0)$. Note however, that if G has a point mass at 0, however small, the risk will be infinite unless $\hat{\lambda}(0) = 0$. This is the only safe decision, since the cannot ascertain that there is no mass at 0.

Note, defining $Z = Y + \mathcal{N}$, $\mathcal{N} \sim Po(h)$ under the KL loss as in Step 1 in the squared loss, might introduce instability due to small values of $\tilde{P}_Z(z-1)$ in the denominator of $\tilde{P}_Z(z)/\tilde{P}_Z(z-1)$, e.g., for $z = \min\{Z_1, ..., Z_n\}$. One might want to define the "corrupted" variable alternatively, as $Z \sim B(Y, p)$. Then $Z \sim Po(p\lambda)$, when $Y \sim Po(\lambda)$. Our smoothing/corrupting parameter is p. We skip the details of the analogouse of steps 1-3.

Throughout the rest of the paper, we consider and evaluate procedures explicitly only under the L_2 loss.

5. Simulations

In this section we provide some simulation results which approximate the risk of various procedures as defined in (1). Specifically for various *fixed* vectors $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_n)$, we estimate $E_{\boldsymbol{\lambda}} \sum \Delta_h(Y_i) - \lambda_i)^2$ and $E_{\boldsymbol{\lambda}} \sum \Delta_{N,h}(Y_i) - \lambda_i)^2$, for various values of h. Our approach is of compound decision and our procedures are permutation invariant, it is known (see, e.g., Greenshtein and Ritov (2009)) that a good benchmark and lower a bound for the risk of our suggested procedures is $nB(\boldsymbol{\lambda})$; here $B(\boldsymbol{\lambda})$ is the Bayes risk for the problem where we observe $\Lambda \sim G$, where G is the empirical distribution which is defined by $\lambda_1, ..., \lambda_n$. The main findings are the following. As already seen in Example 1, adjusting the classical non parametric empirical Bayes yields a significant improvement in the risk. The modification of the normal empirical Bayes works well (for a suitable choice of h), in the Poisson case. It works surprisingly well even for configurations with small λ_i, as may be seen in Table 2.

The class Δ_h of adjusted Poisson EB procedures has advantage over the class $\Delta_{N,h}$ of modified normal EB procedures when applied to configurations with homogeneous λ_i's, see Tables 3 and 4. There are a few reasons for that. First, aggressive smoothing is helpful in a homogeneous setup. Thanks to the bias correction in Adjustment 1 and the Rao-Blackwellization step of Adjustment 2, the adjusted Poisson class performs relatively well when the smoothing is aggressive, i.e., h is large. Another reason is that in a homogeneous setup, much of the risk is due to large and moderate deviations, and the need to correctly assess those few deviations. The normal approximation might be misleading for moderate deviations and thus an advantage to the class of adjusted Poisson procedures might be expected. Another advantage of the adjusted Poisson EB procedures is its little sensitivity to the choice of h, compared to the modified normal procedures. This small sensitivity is evident in the simulations and should be understood better theoretically. Adjustment 2 seems like a crucial stabilizer. Also, the simulations indicate the advantage of the choice $q = \frac{1}{4}$ over the choice $q = 0$.

We elaborate on Table 1. The reading of the other tables is similar. In Table 1 we compare the different estimators when λ_i, $i = 1, \dots, 200$, are evenly spaced between 5 and 15. The table is based on 1000 simulations. We present the risk of Δ_h and $\Delta_{N,h}$ for various values of h. In practice h should be selected by cross-validation or another method, we elaborate on it in Section 6. The normal procedures are based on variance stabilizing transformation with both $q = 0$ and $q = \frac{1}{4}$. Monotonicity is imposed on all the estimators, as described in Step 3, through the Iso-Regression R-procedure . The risk that corresponds to Δ_0 is the risk of the classical Poisson empirical Bayes procedure (i.e., no smoothing through convolution) on which monotonicity is imposed. For the configuration studied in Table 1, the risk of the naive procedures equals the sum of the variances which equals $200 * 7.5 = 1500$. A good proxy of the empirical distribution which is defined by $(\lambda_1, ..., \lambda_{200})$ is $U(5, 15)$. The Bayes risk for the case where $\Lambda \sim U(, 5, 15)$ was computed numerically and it equals approximately 4.4, hence we have $nB(\boldsymbol{\lambda}) \approx 200 \times 4.4 = 880$. In each line the number in boldface

Table 1
Different EB procedures for $\lambda_1, \ldots, \lambda_{200}$ that are evenly spaced between 5 and 15

Δ_h	h	0	0.2	0.4	0.8	1.8	3
	risk	1114	1049	1017	994	**965**	958
$\Delta_{N,h}$	h	0.2	0.3	0.5	0.7	0.9	1.2
$q = 0$	risk	1263	1131	1043	**1022**	1063	1197
$q = \frac{1}{4}$	risk	1230	1099	**1013**	997	1046	1138

Table 2
Different EB procedures for $\lambda_1, \ldots, \lambda_{200}$ that are evenly spaced between 0 and 5

Δ_h	h	0	0.2	1	1.8	2.4	3
	risk	248	**232**	232	242	249	258
$\Delta_{N,h}$	h	0.2	0.3	0.5	0.8	1.0	1.4
$q = 0$	risk	324	291	268	**267**	280	317
$q = \frac{1}{4}$	risk	308	267	245	**242**	254	291

corresponds to the minimal risk.

The model studied in Table 2 is of λ_i, $i = 1, \ldots, 200$ evenly spaced between 0 and 5. Comparing the two halves of the table, one may see how well the normal modification works even for such small value of λ_i.

Next, in Table 3, we study the case where $\lambda_1 = \cdots = \lambda_{200} = 10$. Here the advantage of the adjusted Poisson over the modified normal is clear.

Next we study the following a situation where we have a few large λ_i values: $\lambda_1 = \cdots = \lambda_{200} = 5$, while $\lambda_{201} = \cdots = \lambda_{220} = 15$. There is still a clear advantage of the adjusted Poisson over he modified normal. See Table 4

Finally we investigate a configuration with only $n = 30$ observations spread

Table 3
Different EB procedures for $\lambda_1 = \cdots = \lambda_{200} = 10$.

Δ_h	h	0	0.2	0.4	1	2	3
	risk	253	121	90	54	38	**28**
$\Delta_{N,h}$	h	0.2	0.3	0.5	0.7	0.9	1.3
$q = 0$	risk	369	231	**208**	290	455	822
$q = \frac{1}{4}$	risk	330	197	**180**	265	442	808

Table 4
Different EB procedures for $\lambda_1 = \cdots = \lambda_{200} = 5$, while $\lambda_{201} = \cdots = \lambda_{220} = 15$.

Δ_h	h	0	0.2	0.4	1.2	2.0	3
	risk	665	476	471	**449**	462	483
$\Delta_{N,h}$	h	0.2	0.3	0.5	0.9	1.1	1.4
$q = 0$	risk	857	687	**580**	696	766	854
$q = \frac{1}{4}$	risk	819	613	**550**	653	732	823

Table 5
Different EB procedures for $\lambda_1, \ldots, \lambda_{30}$ that are evenly spread between 0 and 20.

Δ_h	h	0	0.2	0.4	1.2	2.0	3
	risk	867	256	249	256	262	260
$\Delta_{N,h}$	h	0.2	0.3	0.5	0.9	1.2	1.4
$q = 0$	risk	316	303	278	**241**	241	239
$q = \frac{1}{4}$	risk	316	302	280	243	**236**	239

over a larger interval. The λ_i are evenly spread between 0 and 20. For this configuration there is a slight advantage of the modified normal procedure. In order to demonstrate the discontinuity of Δ_h mentioned in Remark 1, we approximated the the risk of Δ_h for $h = 0.01$, based on 1000 simulations. The approximated risk is **244**, compared to 867, for $h = 0$, this is also the minimal approximated risk from the values of h that we tried in Table 5. Note, that 867, the average simulated loss that corresponds to $h = 0$, is much larger than the risk of the naive procedure which estimate λ_i by Y_i; the risk of the later is the sum of the variances which equals $10 \times 30 = 300$.

Remark on sparsity: Our procedure is less efficient in situations with extreme "sparsity". Here by "sparsity" we mean that the vast majority of the values of λ_i are equal to a certain known value λ_0 and very few others are very different from λ_0. In such a cases thresholding procedures might perform better, i.e., "round" to λ_0 the estimators of λ_i for Y_i which are not too far from λ_0.

6. Choosing the smoothing-parameter by Cross-validation

In practice we need to choose h in order to apply our adjusted Poisson method. In this section we will suggest a slightly non-standard way for cross validation. It is explained in the Poisson context, and then in the normal context. The same general idea works for other cases where an observation may be factorized, e.g., for infinitely divisible experiments. About factorization of experiments, see

Greenshtein (1996) and references there.

A standard situation in cross validation analysis is the following. The sample is composed of pairs $Z_i = (X_i, Y_i)$, $i = 1, \dots, n$, where X_i is considered as an explanatory variable, and Y_i is the dependent variable. We have a class of predictors which depends on a parameter: $\{\delta_h(\cdot; Z_1, \dots, Z_n) : \quad h \in H\}$. Often, every value of h represents a different trade-off between variance and bias. The problem is how to select a good value of h. One approach is based on setting aside a test sub-sample, $Z_{m+1}, \dots, Z_n$, say. We can construct the predictors $T_h(\cdot, Z_1, \dots, Z_m)$, $h \in H$, and use the test-bed sample for validation, e.g., compute $S(h) = \sum_{i=m+1}^n \Bigl(T_h(X_i; Z_1, \dots, Z_m) - Y_i\Bigr)^2$. On the face of it, the Poisson sample we consider does not have this structure. There are no explanatory and dependent variables, just one observation, and the observations are not i.i.d., at least not conditionally on $\lambda_1, \dots, \lambda_n$. Yet, we can separate the sample to two independent samples.

Let $p \in (0, 1)$, $p \approx 1$, and let $U_1, \dots, U_n$ be independent given $Y_1, \dots, Y_n$, $U_i \sim B(Y_i, p)$, $i = 1, \dots, n$. As is well known, one of the features of the Poisson distribution is that $U_i \sim Po(p\lambda_i)$, and $V_i \equiv Y_i - U_i \sim Po((1-p)\lambda_i)$, and they are independent given $\lambda_1, \dots, \lambda_n$. We will use the main sub-sample $U_1, \dots, U_n$ for the construction of the family of estimators (parameterized by h), while the auxiliary sub sample, $V_1, \dots, V_n$ for validation. Let $\hat\delta_h^*(\cdot)$, $h \in H$ be a family of estimators, based on $U_1, \dots, U_n$ such that $\hat\delta_h^*(U_i)$ estimates $p\lambda_i$, $i = 1, \dots, n$. Consider:

$$\begin{aligned} &V(h; \boldsymbol{U}, \boldsymbol{V}) \\ &= \frac{1}{n}\sum_{i=1}^n \Bigl(\hat\delta_h^*(U_i) - p(1-p)^{-1}V_i\Bigr)^2 \\ &= \frac{1}{n}\sum_{i=1}^n \Bigl(\bigl(\hat\delta_h^*(U_i) - p\lambda_i\bigr) - p(1-p)^{-1}\bigl(V_i - (1-p)\lambda_i\bigr)\Bigr)^2 \\ &= \frac{1}{n}\sum_{i=1}^n \bigl(\hat\delta_h^*(U_i) - p\lambda_i\bigr)^2 + R_n(h) + A_n, \end{aligned} \tag{12}$$

where A_n is a random quantity that does not depend on h, and has no importance to the selection of h, while

$$R_n(h) = \frac{2p}{(1-p)n}\sum_{i=1}^n \bigl(\hat\delta_h^*(U_i) - p\lambda_i\bigr)\bigl(V_i - (1-p)\lambda_i\bigr). \tag{13}$$

Since $V_1, \dots, V_n$ are independent and independent of $U_1, \dots, U_n$ given $\lambda_1, \dots, \lambda_n$:

$$E(R_n^2(h)|\boldsymbol{U}, \boldsymbol{\lambda}) = \frac{4p^2}{(1-p)n^2}\sum_{i=1}^n \bigl(\hat\delta_h^*(U_i) - p\lambda_i\bigr)^2\lambda_i. \tag{14}$$

We conclude that if $(1-p)n/\max\{\lambda_i\}|H| \to \infty$, then

$$V(h; \boldsymbol{U}, \boldsymbol{V}) = L(\hat\delta_h^*, p\boldsymbol{\lambda}) + o_p(1), \tag{15}$$

uniformly in $h \in H$. Recall that the decision function $\hat{\delta}_h^*$ used in the above result, is the non-parametric empirical Bayes procedure based on $U_1, \dots, U_n$ and $\hat{\delta}_h^*(U_i)$ is estimating $p\lambda_i$. If also $p \to 1$, we suggest to use the value h that minimizes $V(h; \boldsymbol{U}, \boldsymbol{V})$, to construct a similar estimator based on the original sample $Y_1, \dots, Y_n$, estimating $\lambda_1, \dots, \lambda_n$.

$V(h; \boldsymbol{U}, \boldsymbol{V})$, given the sample $Y_1, \dots, Y_n$ is a randomized estimator of the loss function. Once again we suggest in this paper to replace a randomized estimator by its expectation given the sample $E\Big(V(h; \boldsymbol{U}, \boldsymbol{V})\Big|\boldsymbol{Y}\Big)$. This expectation can be estimated by a Monte Carlo integration—sampling K i.i.d. samples of $\boldsymbol{U}$ and $\boldsymbol{V}$.

For the normal model, $Z_i \sim N(\mu_i, 1)$, $i = 1, \dots, n$, let $\epsilon_i \sim N(0,1)$ be auxiliary i.i.d. variables, independent of $Y_1, \dots, Y_n$. Define $U_i = Y_i + \alpha\epsilon_i$, $V_i = Y_i - (1/\alpha)\epsilon_i$. Then U_i and V_i are independent both with mean μ_i, and with variances $1 + \alpha^2$ and $1 + (1/\alpha^2)$ correspondingly. Again, $\boldsymbol{U}$ may be used for estimation and $\boldsymbol{V}$ for validation, where $\alpha > 0$, $\alpha \to 0$.

Example 2: Consider the configuration $\lambda_1 = \cdots = \lambda_{200} = 10$, simulated in Table 3 Section 5. In that table $h = 3$ is recommended with a noticeable advantage over $h \leq 0.4$. We applied the above cross validation procedure with $p = 0.9$ on a single realization of $Y_i,\ i = 1, \dots, 200$. We repeated the cross-validation process $K = 10000$ times on this single realization for the values $h \in \{0, 0.5, 1, 1.5, 2, 2.5, 3\}$. The corresponding numbers (scaled by $(1-p)^2$) were: 165.834, 164.862, 164.736, 164.457, 164.421, 164.286, 164.340. Note that, the last numbers represent mainly the variance of our validation variable, but the success of the corresponding estimator is also a factor. The numbers indicate that the choices $h = 0, 0.5, 1$ are inferior, the formal recommended choice is $h = 2.5$, the second best is $h = 3$.

We repeated the simulation on another single realization, again $K = 10000$, this time we took $p = 0.85$. The corresponding numbers are: 220.562, 217.986, 217.706, 217.374, 217.209, 217.272, 217.247. Again, the numbers indicate that the choices $h = 0, 0.5, 1$ are inferior. The formal recommended choice is $h = 2$, the second best is again $h = 3$.

7. Real Data Example.

In the following we study an example based on real data about car accidents with injuries in 109 towns in Israel in July 2008. The 109 towns are those that had at least one accident with injuries in that period of time, in the following we ignore this selection bias. There were 5 Tuesdays, Wednesdays and Thursdays, in that month. For Town i, let Y_i be the total number of accidents with injuries in those 5 Wednesdays. Similarly, for Town i, let Z_i be the average number of accidents with injuries in the corresponding Tuesdays and Thursdays. We modelled Y_i as independent distributed $Po(\lambda_i)$. In the following we will report on the performance of our empirical-Bayes estimator for various smoothing parameters h. It is evaluated through the 'empirical risk':

$$\sum(\Delta_h(Y_i) - Z_i)^2.$$

Table 6
EB applied to traffic accident by city

Δ_h	h	0	0.5	1	1.5	2	3
	Risk	140	163	172	168	166	159
$\Delta_{N,h}$	h	0.2	0.6	1	2	3	4
	Risk	262	185	174	170	183	202

The towns Tel-Aviv and Jerusalem had a heavy impact on the risk and thus we excluded them from the analysis. The remaining data seems to have relatively low values of λ_i, a case where the classical Poisson-EB procedure is expected to perform well, and indeed it is. The range of Y_i is 0-14, while $\sum Y_i = 135$, and $\sum Y_i^2 = 805$. In this example, the classical Poisson-EB adjusted for monotonicity (i.e., $h = 0$), gave the best result. Applying a smoothing parameter $h > 0$ is slightly inferior based on the above empirical risk. Yet, it is re-assuring to see how stable is the performance of Δ_h, as h varies. The empirical loss for the naive procedure estimating λ_i by Y_i, is 240. The modified normal estimators with $q = \frac{1}{4}$ and various values of h was applied to the data as well. Again a clear advantage of our class of adjusted Poisson procedures over the class of modified normal procedures was observed. In particular, the former class is much more stable with respect to the choice of the smoothing parameter h. The results are summarized in Table 6.

8. The nonparametric MLE

The nonparametric maximum-likelihood (NPMLE) is an alternative approach for estimating δ^G. It yields, automatically, a monotone and smooth decision function. See Jian and Zhang (2009) for the normal model. To simplify the discussion, we will assume that $\lambda_1, \dots, \lambda_n$ are i.i.d. random variables sampled from the distribution G.

Note that the NPMLE maximizes with respect to G, the likelihood function:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}_G(y_i) &= \sum_{i=0}^{\infty} \mathbb{P}_n(i) \log \hat{p}_G(i) \\
&= \sum_{i=0}^{\infty} \bigl(\bar{\mathbb{F}}_n(i-1) - \bar{\mathbb{F}}_n(i)\bigr) \log \hat{p}_G(i) \\
&= \log \hat{p}_G(0) + \sum_{i=0}^{\infty} \bar{\mathbb{F}}_n(i) \log \frac{\hat{p}_G(i+1)}{\hat{p}_G(i)} \\
&= \log \hat{p}_G(0) + \sum_{i=0}^{\infty} \bar{\mathbb{F}}_n(i) \log \hat{\delta}^G(i) + C(\boldsymbol{y}).
\end{aligned}$$

where $\mathbb{P}_n$ is the empirical process, $\mathbb{P}_n(i) = \mathbb{P}_n(\{i\})$, and $\bar{\mathbb{F}}_n(i) = \sum_{j=i+1}^{\infty} \mathbb{P}_n(j)$ ($\bar{\mathbb{F}}_n(-1) = 1$).

Suppose G is supported on $[a, b]$. Extend

$$\delta^G(y) = \frac{\int \lambda^{y+1} e^{-\lambda} dG(\lambda)}{\int \lambda^y e^{-\lambda} dG(\lambda)}, \quad y \in R_+.$$

Then, clearly, $\delta^G(y) \in [a, b]$. Moreover, it is monotone non-decreasing with derivative $\delta^{G'}(y) = \text{cov}(\lambda, \log \lambda) \in [0, b \log b - a \log a]$ (the covariance is with respect to measure $\lambda^y e^{-\lambda} dG(\lambda)$ normalized)

It is well known that the NPMLE is discrete with point mass $g_1, \dots, g_k$ on $\lambda_1, \dots, \lambda_k$ say. It is easy to see that it satisfies

$$\sum_{i=1}^{n} \frac{\lambda_j^{y_i}}{y_i! p_G(y_i)} = e^{\lambda_j} \quad , j = 1, \dots, k.$$

Since the left hand side is a polynomial in λ of degree $\max y_i$, and a polynomial of degree q in λ can be equal to $\exp\{\lambda\}$ only q times, we conclude that $k < \max y_i$ (a more careful argument can reduce the bound on the support size). Hence, it is feasible to approximate algorithmically the NPMLE. Pursuing the asymptotic properties of the NPMLE estimator of $\hat{\delta}^G$ is beyond the scope of this paper and we intend to do it somewhere else.

References

Brown, L. D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. H. (2005). Statistical analysis of a telephone call center: a queing science perspective. *Jour. Amer. Stat. Asoc.* **100** 36-50.

Brown, L.D., Cai, T., Zhang, R., Zhao, L., Zhou, H. The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. To appear *Probability and Related Fields.*

Brown, L.D. and Greenshtein, E. (2009). Non parametric empirical Bayes and compound decision approaches to estimation of high dimensional vector of normal means. *Ann. Stat.* **37**, No 4, 1685-1704.

Clevenson, L. and Zidek, J.V. (1975). Simultaneous estimation of the mean of independent Poisson laws. *Jour. Amer. Stat. Asoc.* **70** 698-705.

Copas, J.B. (1969). Compound decisions and empirical Bayes (with discussion). *JRSSB* **31** 397-425.

Greenshtein, E. (1996). Comparison of sequential experiments. *Ann. Stat.* **24**, No 1, 436-448.

Greenshtein, E. and Ritov, Y. (2008). Asymptotic efficiency of simple decisions for the compound decision problem. To appear in The 3'rd Lehmann Symposium. IMS Lecture Notes Monograph Series, J.Rojo, editor.

W. Hengartner, N. W. (1997). Adaptive Demixing in Poisson Mixture Models. *Ann. of Statist.* **25**, 917–928.

Johnstone, I. (1984). Admissibility, difference equations and recurrence in estimating a Poisson mean. *Ann. Stat.* **12**, 1173-1198.

Johnstone, I. and Lalley, S. (1984). On independent statistical decision problems and products of diffusions. *Z. fur Wahrsch.* **68**, 29-47.

Maritz, J.S. (1969). Empirical Bayes estimation for the Poisson distribution. *Biometrika* **56**, N0.2, 349-359.

Robbins, H. (1951). Asymptotically subminimax solutions of compound decision problems. *Proc.Third Berkeley Symp.* 131-148.

Robbins, H. (1955). An Empirical Bayes approach to statistics. *Proc. Third Berkeley Symp.* 157-164.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann.Math.Stat.* **35**, 1-20. Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

Wenhua, J. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Stat.* **37**, No 4, 1647-1684.

Zhang, C.-H.(2003). Compound decision theory and empirical Bayes methods.(invited paper). *Ann. Stat.* **31** 379-390.